网络学术文档细粒度聚合本体构建研究*

马翠嫦[1] 曹树金[2]
[1] 中山大学图书馆 广州 510275 [2] 中山大学资讯管理学院 广州 510006

摘要:[**目的/意义**]旨在探索网络学术文档细粒度聚合本体构建的理论和方法。[**方法/过程**]在梳理相关理论与方法的基础上,首先明晰细粒度聚合本体概念的基本类型、粒度特征和定义等基本理论问题,然后以网络环境下图书情报学领域“引文分析”主题语料为数据来源,从概念、属性和关系、实例等方面对细粒度聚合单元本体构建进行逐一探讨,并对本体进行评估和讨论。[**结果/结论**]首次提出基于聚合单元知识体系构建细粒度聚合本体的思路与方法,可为基于聚合单元的细粒度组织、检索和导航中知识组织系统工具的构建提供参考。

关键词:信息聚合 本体 网络文档 图书情报领域
分类号:G250
DOI:10.13266/j.issn.0252-3116.2019.24.012

1 研究背景

当前搜索引擎对于海量网络资源的组织仍以网站、网页和各类型文件载体为主要的控制与组织单元,但也出现了对于网页局部信息单元进行搜索和定位的功能,如百度搜索引擎对于百科词条检索结果的揭示可细化到具体知识点。但对于学科领域用户而言,他们对于网络文档的需求往往按照搜索任务情景的不同而分散于不同体裁类型。因此,我们仍然面临如何实现多类型网络学术文档细粒度聚合的问题,即如何让网络信息系统根据用户明确表达的信息需求和检索情景(如:任务、用户偏好等)而为用户呈现经过筛选、抽取和序化的多类型网络资源整体或局部,通过对于聚合单元类型、粒度、关系和属性等方面的控制,来更灵活、准确地为用户呈现所需的目标信息,满足用户对于网络学术资源的需求。

从信息组织的角度而言,网络学术文档细粒度聚合的实现,要求在把握学科领域用户需求的基础上构建领域内共同认可的、反映网络学术资源聚合中多类型网络文档内信息单元的层级、类型及其与用户需求关系的知识组织系统;对该知识组织系统中的知识概念、关系以及概念之间的推理规则进行形式化定义,使得网络资源能通过本体有效地表达出机器能够识别的语义概念,为细粒度聚合单元的抽取、组织、关联与检索匹配提供语义基础。

鉴于此,本文旨在以聚合单元概念为基础探索网络学术文档细粒度聚合本体构建的理论与方法。聚合单元作为细粒度聚合的基本对象,是指以细粒度聚合作为信息组织和检索方式时系统控制和处理的基本文本内容单元,是按照网络学术文档体裁结构划分的不同层级的语言功能单元的统称。网络学术文档的聚合单元既可以是文档整体,也可以是网络资源的局部,如研究论文的结论/讨论部分的段落单元,或结论/讨论部分中的“提出后续研究建议”的句群单元。

围绕研究目标,本文在梳理相关理论和方法的基础上,提出细粒度聚合本体的理论框架,通过实证研究构建细粒度聚合本体,并进行评估和讨论。

2 理论与方法基础

知识单元理论与细粒度聚合本体的构建相关理论密切联系但又有区别:知识单元理论为面向资源载体内信息单元的组织提供理论依据,为基于知识元的知

* 本文系中央高校基本科研业务费项目“支持跨学科知识发现的学术论文信息单元识别与聚合研究”(项目编号:17wkpy56)和国家社会科学基金重大项目“基于特定领域的网络资源知识组织与导航机制研究”(项目编号:12&ZD222)研究成果之一。
作者简介:马翠嫦(ORCID:0000-0002-2478-4714),副研究馆员,博士,E-mail:xx00217@163.com;曹树金(ORCID:0000-0003-1855-4522),教授,博士生导师。
收稿日期:2019-07-30 修回日期:2019-10-28 本文起止页码:107-118 本文责任编辑:易飞

识组织系统构建提供理论与方法基础。然而,由于知识单元理论中定义的知识组织对象仅为包含知识单元的信息内容,而细粒度聚合本体中定义的知识组织对象是按照语言功能进行定义的聚合单元,聚合单元的组织不仅涉及知识单元组织的问题,还涉及了这些聚合单元的语言功能带来的语义关系及其与用户搜寻形成的关联关系,因此知识元理论未能涵盖细粒度聚合所需的知识组织系统构建的全部问题,需要结合聚合单元构建的理论基础——体裁理论与方法,建立聚合单元知识体系构建的理论与方法基础。此外,细粒度聚合本体的构建离不开本体的基本理论与方法。因此,本研究梳理知识单元相关理论、体裁结构规则相关理论与本体构建理论,从而构建适应细粒度聚合本体构建需求的理论框架。

2.1 知识单元相关理论研究

针对文献内部知识单元的研究曾受到各学科领域的关注,但学界对于知识单元的认识尚未统一。知识单元概念在不同时期和不同学科领域研究中的定义也各不相同,可包括"知识基因""知识概念""知识节点""知识因子""知识点""知识元""知识链接""知识单元"等,但都是指一定单元内具有独立含义的知识内容,既可以是文献,也可以是文献的局部片段,还可以是文献中包含的概念等知识要点[1]。对于知识单元的研究,可以分为面向知识组织的学科领域知识单元研究、面向知识抽取与利用的知识单元研究和教育领域面向课程的知识元组织研究:

2.1.1 面向知识组织的学科领域知识单元研究　根据文庭孝的观点,知识单元按照知识组织的发展阶段和深入程度可分成文献单元、信息单元和知识单元等 3 种主要的形态。其中知识单元的实践与研究始于文献单元。文献单元作为天然的、包含知识的载体单元,自然地成为知识管理的初始单元,并在此基础上逐渐形成了完善的知识体系。因而,文献单元作为知识单元的早期形态,其中包含着知识单元,而知识单元最终附着在一定形式的文献单元中,体现为文献单元[1]。

相关研究强调知识组织的对象应深入至文献内知识单元层,因而弥补了既有知识组织理论对文献内容反映不足的缺憾,但对于如何划分文献内部的知识单元的粒度和层级仍然未有进一步的理论和方法指引,因而未能满足细粒度聚合知识组织系统对于聚合单元划分的要求。

2.1.2 面向知识抽取与利用的知识单元研究　温有奎等从知识组织的角度系统地提出了"知识元"理论:"假定文本内容的组织排列是由一个个独立知识元素的逻辑排序结构,这种独立的知识元素称为知识元,逻辑依存关系称为知识链。知识元是构造知识结构的基元。""知识元及其结构组成不同的知识单元。"[2]。知识元的类型包括:描述型(信息报道、名词解释、数值、问题描述、文献引证等)和过程型(步骤、方法、定义、原理、经验等)两大类[2]。温有奎等对于知识元理论及其知识组织方法进行了系统研究[2-9],在此基础上 CNKI 构建了学术论文中定义、数字和图表等类型知识元的搜索系统[10]。

可见,知识元理论中的知识元概念始于知识组织对象粒度的细化,这与本研究细粒度聚合本体中包含的聚合单元概念类似却又有所不同。类似的是组织对象粒度从文本深入到文本内容,不同的是知识元理论对于知识元类型(如定义、数值、图表等)的划分着眼于知识的组织与利用,旨在构建基于知识实例及其关系的知识库;而本研究中对于聚合单元类型的划分则着眼于有用信息片段的组织与利用,旨在构建信息片段之间及信息片段与用户任务情景之间的关联关系。因而,知识元定义、抽取、本体构建和组织的相关研究,可为聚合单元的抽取和组织提供众多的方法基础,也为聚合单元本体构建提供参考。

2.1.3 面向课程的知识元组织研究　此方面研究主要集中在教育技术领域,该领域中知识元常被称为"知识点",即由不同的知识点根据其相关性组成知识体系。知识点是教学活动过程中传递教学信息的基本单元,包括理论、原理、概念、定义、范例和结论等。知识点可进一步分解,在结构上不可分割的知识点称为原子知识点。相关的一组知识点集成为知识单元。知识点划分的基本原则是保证知识内容的局部完整性,而其大小可随需要而定,可能相差很悬殊。例如,一章可划为一个大的知识点,其中一节的内容又可细划为较小的知识点,一节中的定义、定理等还可以划为更小的知识点。有学者以教育技术学科领域的知识分类体系为基础构建以知识元为单位的教育技术学科资源库[11]。此外,有学者在知识元理论的基础上提出面向知识组织与共享的教育资源知识元描述模型,探索知识摘要和知识融合的相关方法和技术[12]等。

教育学领域知识元研究构建的知识体系实际上以课程知识组织与教育为主要目标,而并非面向资源的组织,因此其知识元实例是真正的知识本身,而非这个知识概念对应的资源。然而,关于知识元本体构建的方法和技术,可以为基于聚合单元知识体系的本体概

念与关系构建提供参考。

2.2 体裁结构规则及其知识体系研究

体裁按照其交际目标而呈现一定的形式特征和结构规则。虽然大部分关于体裁理论的研究都采用典型体裁，如科学论文、新闻或短篇小说作为研究案例，但现在普遍认为体裁结构存在于所有交流网络中，而大部分专业的交流通常无需依靠外部联系，而是依赖于体裁结构来完成其共同的工作[13]。关于体裁结构规则的研究可分为语言学领域对 Swales 模型的探索和发展、图书情报领域对于体裁结构的利用与探索这两方面进行总结。

2.2.1 语言学领域对 Swales 模型的探索和发展　体裁结构研究代表性理论可包括 J M. Swales 的学术论文"语轮－语步"分析模型。该模型在研究论文体裁所特有的介绍－方法－结果－讨论的构成基础上，按照研究论文的目标进一步对"介绍"部分的内容进行语轮－语步分析，从而将研究论文划分成由构成(component)－语轮(move)－语步(step)不同粒度层级组成的信息单元[14]。

在 J M. Swales 的初始模型的基础上，众多学者将体裁分析理论和方法用于自然科学、生物医学、社会科学[14－16]、野生生物行为研究和生态保护领域[17]等进行检验。Swales 模型还被拓展至研究论文的其他构成单元进行研究，如对于摘要的研究[18]、对于方法的研究[19]、对于结果的研究[20]、对讨论的研究[21]和对所有构成的研究[21－22]。B A. Lewin 等对社会科学领域语料的导言和讨论部分的语轮和语步进行了全面研究[23]，检验和丰富了研究论文的体裁结构知识体系。

此外，国内学者也对语篇的体裁结构进行了研究，如：赵福利参考 Bhatja 的语轮模式，研究电视新闻导言的语轮结构[24]；葛冬梅和杨瑞英参考 Bhatja 的语轮模式，对学术论文的摘要进行研究[25]；偓艳嫣和王同顺参考 Swales 模型对英语学术讲座的结构进行研究[26]；杨瑞英参考 Swales 模型对英语语言功能学的学术论文各构成进行语轮和语步分析，并提出了理论研究类学术论文的构成、语轮和语步[27]。

2.2.2 图书情报领域对于体裁结构的利用与探索　自从上世纪 90 年代末数字图书馆项目出现以来，数字文档的解构与重组、数字资源信息单元的识别与利用问题开始受到学界的关注[28]。这些关于数字文档划分与利用的研究，大都以体裁结构理论为基础，结合用户的信息获取任务进行划分和关联分析的。

如 A. Dillon 检验了人们对于网络新闻这一体裁内容结构的认知，从而为基于体裁的理解与文内导航提供了依据[29]。A. Dillon 及其同事还从用户认知的角度围绕语篇的逻辑结构和语义结构对用户迷航问题和导航需求进行了系列研究[29－34]。L. Zhang 构建了心理学期刊论文的功能单元分类体系，识别出期刊文章组成(例如：介绍、方法、结果和讨论)中的最小信息单元[35－36]，并对阅读信息获取的效率和效用进行了检验[37]。C-C. Ma 和 S-J. Cao 构建了图书情报学领域跨体裁的聚合单元分类体系[38]。

从已有研究可以看到，面向用户认知与需求的体裁结构划分，对于提高信息利用的效率与效用有重要作用，可为细粒度聚合提供理论支持。更重要的是，无论是语言学领域还是图书情报学领域对于体裁结构的划分与利用研究，都形成了一系列关于体裁结构下语言功能单元的知识体系。然而，目前关于信息单元利用的研究尚处于探索阶段，未进一步从知识组织的角度进行考量，更未构建相应的知识组织系统以支持实际的应用。

2.3 本体构建的理论与方法

本体研究既有的理论与方法，可为细粒度聚合本体构建提供直接的理论与方法基础。以下从本体的类型、构建原则与方法、构建工具、评估方法等方面进行梳理：

2.3.1 本体是共享概念模型明确的形式化规范说明　目前已有大量关于本体的研究，特别是国外，众多的研究组织和机构根据各自需求建立了多类型的本体。按照应用范围和层次进行划分，本体可分为通用本体、领域本体和应用本体。通用本体不针对具体的领域知识，可进行跨领域范围的复用；领域本体则表达特定学科领域的知识体系；应用本体是为特定应用而创建的本体知识库，可包括跨学科领域的知识。其中，通用本体和领域本体是应用本体的上层本体[39－41]。

更具体地，R. Mizoguchi 等提出按照本体的应用目的进行划分，分为领域本体、顶级(通用)本体和任务本体 3 种，其中任务本体是指通过顶层概念表达具体任务专用的概念类、属性和关系，它描述特定任务或行为中的概念体系，提供可解答与某具体任务或行为有关问题的概念集[41]。N. Guarino 提出按照本体概念的具体程度和本体概念相对于领域的独立性进行类型划分，其中，按照本体概念的详细程度分为较详细的参考本体(Reference ontology)和较简略的共享本体(Share ontology)，按照本体概念对学科领域的独立性可划分为 4 类：顶级本体(Top-level Ontologies)、领域本

体(Domain Ontologies)、任务本体(Task Ontologies)和应用本体(Application Ontologies)[42-43]。此外,有学者按照本体应用将其划分为领域本体、通用本体(常识本体)、知识本体、语言学本体和任务本体等5种类型[44-45]。

2.3.2 关于本体构建的理论和方法较为成熟 其中,较为典型的本体构建原则是T. Gruber提出的本体构建五原则,即清晰度、一致性、可扩展性、中立性和最小本体承诺[46]。而较为典型的构建方法包括:骨架法、IDEF5法、七步法、五步循环法、METHONTOLOGY法、TOVE法、KACTUS法、SENSUS法和循环获取法等[45]。

2.3.3 Protégé是本体构建的重要工具 Protégé具有可视的用户界面,支持DAML + OIL和OWL语言,可实现模块化设计[47],且可利用本体描述语言进行系统外的修改。由于Protégé具有开源代码、有中文版本等诸多优点,在国内被广泛采用[47]。

2.3.4 本体评估的方法较为多样 包括用户评价法、应用评价法、语料库评价法、专家评价法和复合指标评价法、黄金标准评价法等。这些评价方法都有其适用性和可操作性,但有研究指出这些本体评价的方法本身也存在一定的局限性,跨领域的适用性并不理想,难以大规模应用。当前,构建指标体系是最为常见的评价方法[47]。此外,也有学者指出,无论在国内还是国外,本体评价方法的研究尚处于探索阶段,缺乏被广泛认可的评价理论体系和评价方法体系,评价集中于概念、属性、关系等方面,仍未创建出综合的本体评价体系,也未出现权威的评价标准[47]。

3 细粒度聚合本体的构建

3.1 细粒度聚合本体的理论框架

3.1.1 细粒度聚合本体概念的基本类型 本体构建的过程一般包含两个阶段,第一阶段目标是确定本体的概念集合,建立核心术语集合;第二阶段目标是确定概念之间的关系。根据网络资源细粒度聚合目标和框架,本文将细粒度聚合本体的基本概念定义为4类,即网络文档、聚合单元、学科领域概念和任务情景,如图1所示:

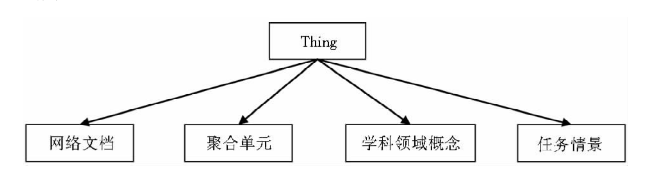


图1 细粒度聚合本体的4类基本概念

在此基础上,细粒度聚合本体构建的思路是:①通过聚合单元知识体系确定聚合单元概念集合;②通过学科领域知识组织系统构建的方法确定学科领域概念集合;③通过已有研究确定任务情景概念集合。由于学科领域本体构建已有大量研究,而任务情景本体又较为简单,本文着重探索基于聚合单元知识体系的本体构建方法。

3.1.2 细粒度聚合本体的粒度特征 从知识组织的已有理论可知,知识组织体系细致程度对网络资源检索和利用效率会产生影响,知识粒度越细则描述的准确性越高,信息片段越小则检索的相关性越高。因此,本文按照网络资源细粒度聚合本体框架中提供语义知识的两个重要知识库——学科领域知识体系和聚合单元知识体系,将两组网络资源知识组织系统中粒度特征分为学科领域概念粒度和受控单元粒度进行定义,如表1所示:

表1 网络资源知识组织系统粒度层级

层级	学科领域概念粒度	层级	聚合单元粒度
K1	主要概念	C1	体裁类型
K2	子概念	C2	构成单元
Kn	…….	C3	语轮-语步单元

学科领域知识体系与聚合单元知识体系粒度的明晰,有利于从领域知识的准确性和网络学术文档的准确与相关性两方面增加网络资源细粒度聚合的效率与效用。

3.1.3 细粒度聚合本体概念的定义 对于细粒度聚合本体而言,由于前期研究已形成聚合单元知识体系[38],已有研究也对任务和任务与信息单元的关联性进行调查[36,49],因此,其核心术语及其属性较容易确定,均可定性地实现形式化说明。

本文将学科领域概念C定义为一个四元组:

$C = \{C_0, P_C, R_C, Syn_C\}$

其中,C_0代表学科领域概念;P_C代表该学科领域概念的一般属性;R_C表示学科领域概念关系集合;Syn_C表示学科领域概念C_0的同义词集合。

本文将任务情景T定义为一个三元组:

$T = \{T_0, C_T, P_T, R_T\}$

其中,T_0代表任务类型的概念;C_T代表该任务主题对应的学科领域概念;P_T代表该任务的一般属性;R_C表示任务概念关系集合。

本文将聚合单元概念A的完整概念定义为一个五元组:

$A = \{A_0, C_A, U(A_0, T), R_A, Syn_A\}$

马翠嫦，曹树金. 网络学术文档细粒度聚合本体构建研究[J]. 图书情报工作,2019,63(24):107－118.

其中，A_0 代表聚合单元的概念；C_A 代表该聚合单元主题对应的学科领域概念；$U(A_0, T_0)$ 表示聚合单元的任务情景相关性，该属性由任务的类型 T_0 及其与该聚合单元 A_0 的相关性程度决定。例如：聚合单元“网络百科”具有百科类体裁的属性，属于体裁层级，其语义功能是介绍相应概念各方面知识，在“学习背景”任务下具有较高的感知可用性。U 为实例下感知有用性的具体数值；R_A 表示关系集合；Syn_A 表示概念 A_0 的同义词集合。

本文将网络文档概念 D 的完整概念定义为一个九元组：

$D = \{D_0, C_D, Tit, Cont, Auth, Inst, S, G, Time\}$

其中，D_0 代表网络文档的概念；C_D 代表该聚合单元主题对应的学科领域概念；Tit 代表文档题名；Cont 代表文档内容，Auth 代表文档作者；Inst 代表文档机构；S 代表文档来源，G 代表文档体裁，Time 代表出版时间。

3.2 细粒度聚合本体的构建与形式化

本文以前期研究构建的图书情报领域语料库中“引文分析”主题的 81 种网络文档为数据来源[38]，通过实证研究的方法建立细粒度聚合本体的 4 类概念集合、属性、关系及相应的实例，构建本体并实现形式化。实验语料所包含的体裁包括开放获取研究论文、在线题录、网络百科词条和学术博文 4 种类型。

3.2.1 细粒度聚合本体概念体系　按照细粒度聚合本体概念的基本类型，其概念体系包括：聚合单元概念体系、学科领域概念体系、任务概念和文档概念 4 个部分。

（1）聚合单元概念体系。按照聚合单元知识体系构建聚合单元的概念集合及概念间关系[38]，采取自上而下的顺序进一步确定聚合单元本体的概念及其属性：根据聚合单元分类体系确定不同层级的聚合单元概念，如研究论文、导言、介绍论题背景等不同层级的概念，如表 2 所示：

表 2　聚合单元知识体系

一级类	二级类	三级类	一级类	二级类	三级类
OA 论文/在线题录	摘要	概述论题	网络百科词条	词条摘要	词条概述
		概述方法			词条基本信息点
		概述结果		词条简介	定性叙述
		概述结论			介绍历史沿革
OA 论文	导言	建立一个领域论题			介绍基本事实
		回顾已有研究			叙述生平事迹
		评述已有研究		知识要点	参阅资料
		呈现当前研究			关键信息点
		厘清定义		人物影响	介绍主要成就
	理论背景	提出理论或概念			列举所获荣誉
		评述理论			列举主要作品或观点
		将理论与当前研究联系起来			介绍相关评价
	论证	介绍论题背景	博客文章	引入	建立交流性论题
		提出作者立场			插入题外交流性话题
		论述理论立场			介绍论题相关客观信息
		作出合理论断			记载论题相关事件
	方法/数据	介绍方法背景		观点	提出博主观点和立场
		描述数据			论述博主观点和立场
		介绍分析方法与程序			总结博主观点和立场
		厘清定义		来源和链接	列出相关链接
	结果	下文提要			列出资源来源
		介绍结果背景		交互评论	提问与回复
		描述所开展的分析			评论与回复
		证明方法或程序合理性			
		呈现结果			
		评论结果			
	讨论/结论	总结研究背景和研究概况			
		下文提要			
		总结结果、结论或理论观点			
		讨论结果			
		评价结果			
		后续研究建议			

(2)学科领域本体构建。学科领域资源本体构建的研究已有较长历史,具有较为成熟的理论和方法基础。为配合探索细粒度聚合本体的构建方法,本文采用基于词典的机器辅助的方式构建主题概念知识体系,具体过程是:以百度百科中"引文分析"词条中关于引文分析的知识体系作为"引文分析"语料库知识体系的基础,采用武汉大学开发的 ROST CM 软件对语料库文本进行分词和词频计算,获取反映学科领域特征的高频关键词概念,经过课题组成员逐一讨论,将全部有意义的新词添加到引文分析知识体系进行完善,从而采用"自上往下"法构建"引文分析"本体。最终构建包括 6 个层级和 100 个概念的学科领域概念体系。

(3)网络文档本体与聚合单元本体。网络文档本体旨在构建关于网络文档各维度信息单元的本体概念及其概念关系,从而与聚合单元概念体系配合,支持网络文档细粒度聚合。本文在参考朱嘉贤等关于 Web 资源本体和邱均平等关于馆藏资源语义本体研究的基础上[40, 50],构建网络本体文档的主要概念。因而,文档本体包括体裁、内容、创作者、单位机构、来源、题名等概念,见图 2。

任务情景概念集定义关于任务情景的概念,从而为构建任务与聚合单元概念之间的关联关系提供基础,支持网络文档细粒度聚合。本文在 L. Zhang 和 L. Freund 等关于任务本体定义的基础上[36, 49],构建任务本体的主要概念,见图 3。

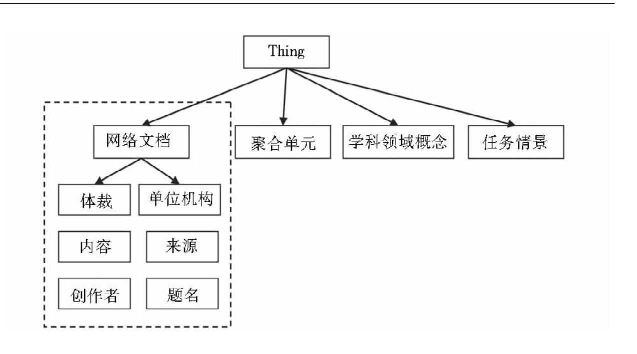


图 2　细粒度聚合本体网络文档本体的概念

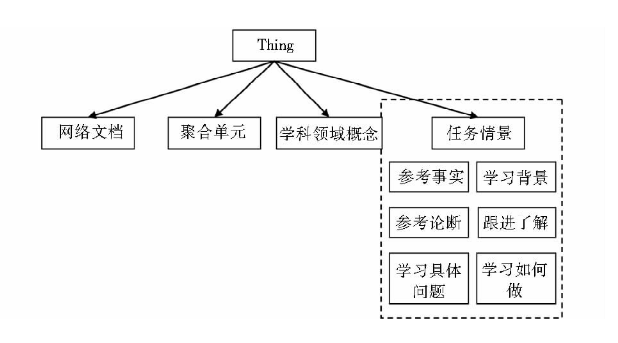


图 3　细粒度聚合本体任务情景本体的概念

3.2.2　细粒度聚合本体属性及关系　在前一阶段析出各类细粒度聚合各类本体概念集合的基础上,第二阶段将明晰本体概念的主要属性,可包括:聚合单元的感知有用性,聚合单元特有的语言功能形成的聚合单元类与类之间、实体与实体之间的语义关系,网络文档本体概念之间存在的关系,以及主题与聚合单元本体和任务本体之间的关系。因此,细粒度聚合本体主要包含 11 种主要属性,如表 3 所示:

表 3　细粒度本体概念属性的定义

类	属性	定义
聚合单元	任务下的感知有用性(数值属性)	是指特定任务下聚合单元的感知有用性
聚合单元	语义推进关系/逆关系	是指特定交际目标下,同组内不同聚合单元语义功能之间所形成的语义推进关系
聚合单元	属于/包含______	是指下级聚合单元与其所属的上级聚合单元之间的属于关系及其逆关系
聚合单元/任务情景	主题是________	是聚合单元或任务情景所包含的主题是…
学科领域概念	是______的学科领域概念	是___________概念的学科领域概念
作者	来自________机构	是指作者来自….机构
作者	撰写________网络文档	是指作者撰写了…..网络文档
机构	包含________作者	是指某机构包含….作者
内容	由________撰写	是指网络文档由….撰写
内容	主题是________	是指该网络文档的内容包含…主题
内容	创作时间是________	是指该网络文档内容的创作时间是…

依据细粒度本体类与属性之间的关系,可以进行如下推理:

(1)聚合单元的有用性。如果某任务情景下某聚合单元"感知可用性"得分高于阈值,则该聚合单元在该任务情景下具有较高的有用性。

(2)聚合单元的任务相关关系。如果特定任务情景下,某些聚合单元的感知可用性得分均高于阈值,则这些聚合单元存在任务相关性。

(3)聚合单元的主题。下级聚合单元继承上级聚合单元“学科领域概念是…”的属性。如果上级聚合单元的学科领域概念是…则下级聚合单元的学科领域概念也是…. . 上下级聚合单元之间存在等价的逆关系。

(4)聚合单元的文档本体相关属性。各层级聚合单元均获得文档本体相关的属性如果文档的作者/机构/来源/题名/体裁是…则该聚合单元的作者/机构/来源/题名/体裁是…文档本体与聚合单元之间存在等价的逆关系。

根据表2及基于类属性的语义推理形式化描述，本文对这些语义关系类型进行了明确形式化说明，即自顶向下地确定了细粒度聚合本体主要包含的5类主要关系：

(1)继承关系及其逆关系。包含子类对父类的继承关系、实例对类的继承关系、类与属性之间的关系等及其逆关系。聚合单元本体中概念之间的语义关系比较明确，可依据聚合单元知识体系中上下位聚合单元之间的关系确立；对于继承关系的属性，可按照关系的来源、性质等方面进行定义；类所包含的实例之间的语义关系，可通过所属类之间的关系获得。

(2)推进关系及其逆关系。聚合单元本体中相同父类的子类及其实例之间的语义推进关系及其逆关系，可依据聚合单元知识体系中上下位聚合单元之间的关系确立。

(3)任务相关性。由于特定任务类型下特定聚合单元具有较高的感知可用性，因而相同层级或不同层级聚合单元之间形成基于任务情景的关联关系，除了可以进行定性说明外，还可以通过聚合单元感知有用性的数值关系进行计算并进行定量的形式化说明。

(4)学科领域概念的语义关系。由学科领域概念之间形成的语义关联关系；相同层级或不同层级类所包含的实例之间的语义相关度则可在领域本体的概念相关度的基础上结合本类聚合单元的情景关联属性进行加权计算。

(5)网络文档概念中的关系。网络文档及其作者、机构等之间的关联关系，可从网络文档元数据中直接获取，也可从机构网站等公开信息源获取和整合。

3.2.3 细粒度聚合本体的实例 以“引文分析”语料库语料作为细粒度聚合本体实例的来源。按照细粒度聚合本体对语料库中网络文档的相关实例信息进行提取与统计，得出本体类的实例数量分布如表4和表5所示：

表4 引文分析聚合单元本体类的实例统计

类型	体裁单元	构成单元	语言功能单元
OA期刊论文	28	136	805
在线文摘	18	30	21
百科词条	13	36	123
学术博文	22	52	298
合计	81	254	1 247

表5 引文分析文档本体及学科领域概念本体类的实例统计

类型	数量	构成单元	数量
学科领域概念	100	机构	50
体裁	4	来源	38
内容	81	题名	81
作者	89		

其中，表4是关于聚合单元知识体系中本体类的实例统计，表5是文档本体和学科领域概念本体类的实例统计。从表4和表5可知，聚合单元本体实例来源于4类体裁的81个实例，体裁单元下所包含的构成单元共计254个，构成单元所包含的语言功能单元共计1247个。学科领域概念本体中包括实例100个，文档本体包含作者89个、机构50个、来源38个。

对于细粒度聚合本体的类而言，其实例的属性可按照本体中该类的属性进行定义。在此着重探索聚合单元实例基于计算而获得的数值属性–聚合单元实例的感知有用性。在聚合单元属性定义阶段，感知可用性作为聚合单元类的数值属性，受聚合单元类型及用户任务情景的影响，是聚合单元任务相关性的提示器。

3.2.4 基于Protégé的细粒度聚合本体形式化 在明晰细粒度聚合本体概念及其属性的基础上，对其进行形式化说明，从而形成形式化的本体。以“引文分析”数据集为语料来源，采用本体编辑和可视化工具Protégé对细粒度聚合本体按照OWL语言规范添加语义标记，进行编码、形式化，从而建立网络文档细粒度聚合本体，其全貌见图4。

从构成本体的大类来看，细粒度聚合本体包含聚合单元本体、学科领域本体、网络文档本体、任务本体，因此，按照层级结构关系查看最上层本体概念，可得到细粒度聚合本体的主体构成及其关系，见图5。

从图5可见，聚合单元本体是细粒度聚合本体的基础，为资源的细粒度聚合提供层级关系、语义关系及任务相关关系，从而支持基于聚合单元间关系的网络文档细粒度聚合和可视化的导航与检索方式。聚合单元本体的层级概貌如图6所示：

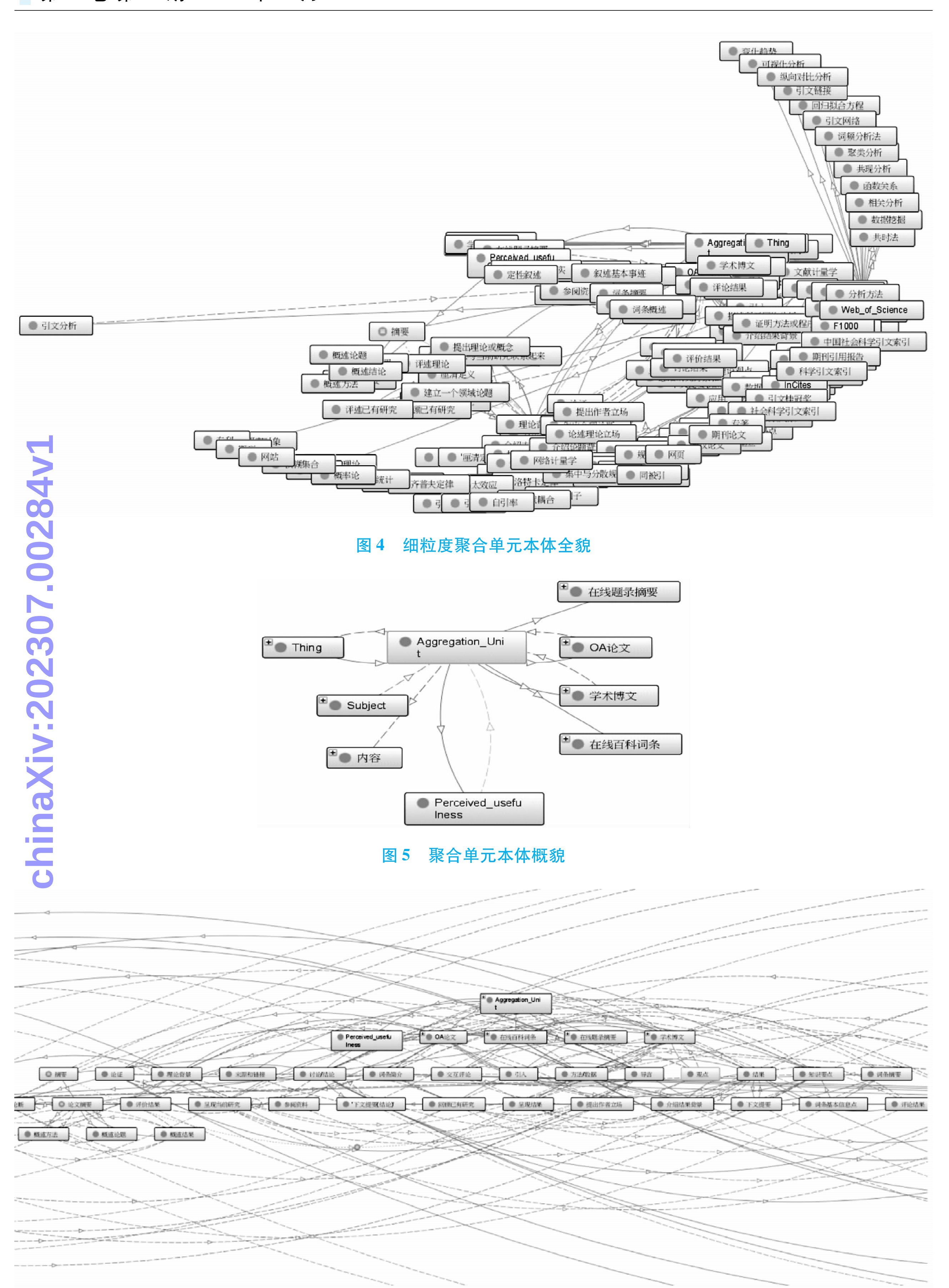


图 4　细粒度聚合单元本体全貌

图 5　聚合单元本体概貌

图 6　聚合单元本体的层级结构(局部)

马翠嫦，曹树金. 网络学术文档细粒度聚合本体构建研究[J]. 图书情报工作,2019,63(24):107－118.

从图 6 可见,聚合单元属于文档本体中内容属性的一部分,包含感知可用性的属性,被学科领域概念描述,包含在线题录摘要、OA 论文、学术博文、在线百科词条等类型的网络文档体裁。各层级聚合单元之间存在整体与部分关系,同一层级下的同组聚合单元之间存在推进关系及其逆关系。

基于“引文分析”网络文档语料库的资源学科领域本体的概念关系层级结构如图 7 所示:

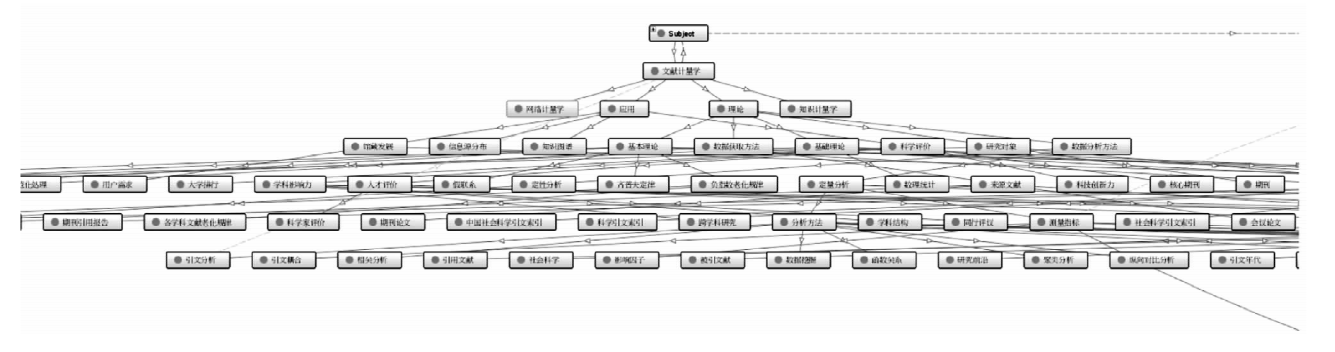


图 7　学科领域概念本体的层级结构概貌(部分)

文档本体与任务本体及其与学科领域和聚合单元之间的关系见图 8。由此可见,学科领域概念是文档本体、任务本体与聚合单元本体的属性。聚合单元来源于文档本体的内容属性。

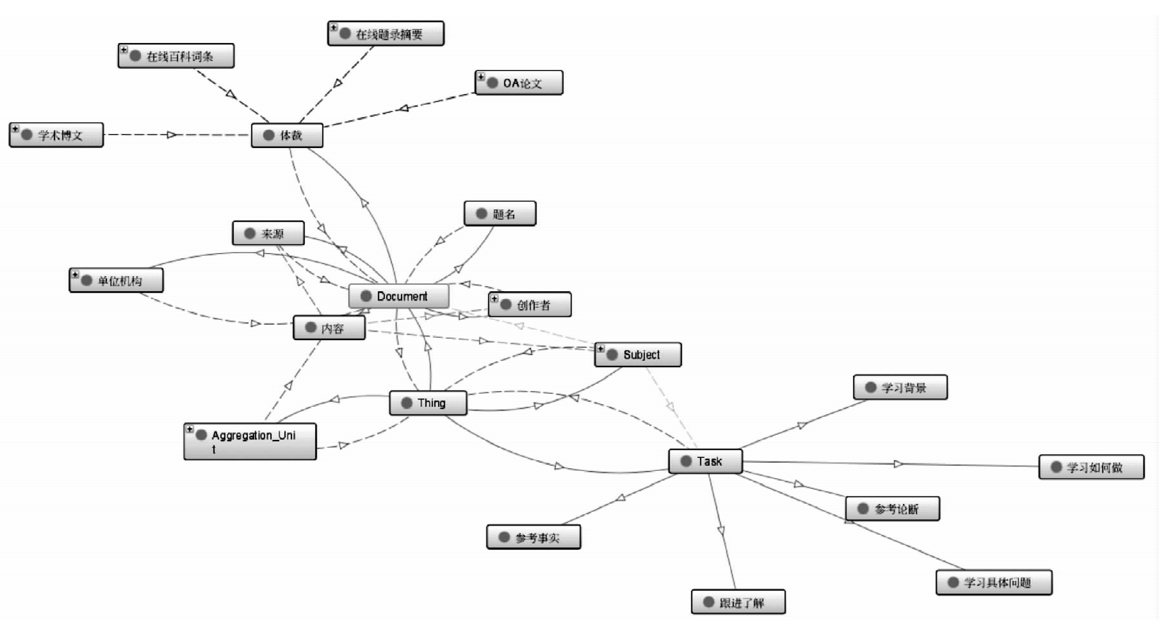


图 8　文档本体与任务本体概貌

4　细粒度聚合本体评估与讨论

岳丽欣和刘文云综合国外各种评价指标提出完整性、清晰性、一致性、可扩展性和兼容性的本体评价标准[51],本文围绕这几个标准方面对细粒度聚合本体进行讨论:

4.1　完整性方面

由于所构建的细粒度聚合本体源于实验语料,因而可较大程度地覆盖语料涵盖各类本体的概念及其关系,尤其是文档本体的覆盖程度达到 100%。与 C-C. Ma 和 S-J. Cao 所划分的聚合单元初始分类体系相比[38],由于本研究提出的聚合单元本体并未采用初始分类体系中包含的两种评分较低的语义功能单元,因此聚合单元本体的覆盖程度为 96.5%。

然而,由于本研究提出的任务本体概念来源于 L. Zhang 和 L. Freund 研究提出的任务类型[36,49],其完整性和系统性与 Y. Li 研究提出的任务分面分类体系相比存在不足,因此后续研究可以参考 Y. Li 提出的任务体系[52],从更多的分面和类型构建任务与聚合单元之间的关联关系,建立更完整的任务本体。此外,对于图情领域更多体裁类型的网络文档甚至是更多学科领域的网络文档而言,本研究提出的细粒度聚合本体不

仅需要补充和完善聚合单元的类型和关系,还应基于词典或基于大规模语料构建学科领域概念体系,从而确保细粒度聚合本体的完整性。

4.2 清晰性方面

细粒度聚合本体的四类概念均可按照既有知识体系构建成层次清晰、概念边界明确、属性和关系定义明确的本体。其中,聚合单元知识体系按照语言功能学中的体裁结构理论构建而成[38],因而聚合单元在概念、属性和关系方面都具有较为明确的定义;任务本体参照 L. Zhang 和 L. Freund 所采用的任务定义明晰不同任务的含义与属性[36, 49];文档本体的概念更被普遍认知;而学科领域概念则在自动分词的基础上,参考网络百科词条并通过图情领域研究人员共同确定,因而其概念含义和概念间关系具有明确的定义,从而保证细粒度聚合本体的清晰性。

4.3 一致性方面

由于作为主体的聚合单元概念的类、属性和关系的数量远远少于学科领域概念体系,且聚合单元知识体体系通过人工内容分析和语轮 – 语步分析的方式产生,分类体系的构建本身就经过对于所划分的聚合单元的内部一致性调查[38],因而并不存在半自动化/自动化构建过程产生的噪音数据,一致性程度较高。

4.4 可扩展性方面

本研究构建的细粒度聚合本体虽然尚处于方法探索人工构建的阶段,但所构建的基本框架,尤其是基于体裁结构理论的聚合单元概念体系允许维护更新以实现本体的进化,可根据实际情况采取自动化或半自动的方法不断完善层次结构和语义,扩充新出现的术语、概念以及关系。

4.5 兼容性方面

细粒度聚合本体包含的聚合单元知识体系、任务知识体系因有统一而明确的理论基础,可实现多个学科领域聚合单元知识体系和任务体系间的兼容;而在百科词条的基础上结合语料词频所选出的学科领域概念使得概念体系具有兼容和映射的基础。

5 小结

本研究以网络学术文档细粒度聚合本体构建为目标,在厘清细粒度聚合本体理论框架的基础上建立细粒度聚合本体并进行评估。首次提出基于聚合单元知识体系的细粒度聚合本体构建的思路和方法,明晰了细粒度聚合本体概念的基本类型,厘清聚合单元粒度与学科概念粒度的关系,对本体概念进行定义,从而构建了细粒度聚合本体构建的理论框架。此外,从本体概念体系、属性及关系、实例构建 3 个方面明晰了本体构建的思路和方法。

本研究构建了面向图书情报学领域 4 种体裁网络文档和用户需求的细粒度聚合本体。然而,由于多种体裁的聚合单元划分尚处于探索阶段、划分难度较大、耗费时间长、又尚未开发出稳定的自动分类方法,本研究采用以人工划分方式为主的小规模语料样本作为数据来源。虽然语料数量较小为细粒度聚合本体的应用带来了局限,但却能在确保概念体系准确的前提下进行构建方法和本体效用的探索性研究。

此次的实验语料样本及相应本体将应用到细粒度聚合原型系统中,进一步对各类聚合单元的自动分类、组织与索引、交互研究等方面进行系统探索,以期更全面地探索和把握聚合单元的效用,为更大规模的、自动化、智能化探索与应用提供基础。

参考文献:

[1] 文庭孝,罗贤春,刘晓英,等. 知识单元研究述评[J]. 中国图书馆学报,2011, 37(5):75 –86.

[2] 温有奎. 焦玉英. 基于知识元的知识发现[M]. 西安:西安电子科技大学出版社, 2011.

[3] 温有奎. 基于“知识元”的知识组织与检索[J]. 计算机工程与应用, 2005, 41(1): 55 –57,91.

[4] 温有奎,徐国华. 知识元链接理论[J]. 情报学报, 2003, 22(6): 665 –670.

[5] 温有奎,温浩,徐端颐,等. 基于知识元的文本知识标引[J]. 情报学报, 2006, 25(3): 282 –288.

[6] 王燕,温有奎. 文本单元向知识单元转化的研究[J]. 情报理论与实践, 2007, 30(3): 409 –411,362.

[7] 温有奎,焦玉英. 基于范畴论的知识单元组织与检索研究[J]. 情报学报, 2010, 29(3): 387 –392.

[8] 温有奎,焦玉英. Wiki 知识元语义图研究[J]. 情报学报, 2009, 28(6): 870 –876.

[9] 温有奎,焦玉英. 知识元语义链接模型研究[J]. 图书情报工作, 2010, 54(12): 27 –31.

[10] 周秀会. 知识元搜索引擎:CNKI 知识搜索平台[J]. 现代情报, 2007, 27(5):220 –222.

[11] 陶善菊,刘清堂,王凡,等. 基于知识元的教育技术学科资源库构建[J]. 现代教育技术, 2011, 21(5): 115 –120.

[12] ZOU J, LIU Q. A knowledge element model for knowledge abstract and fusion system[C]//2009 International conference on new trends in information and service science. Washington, DC: IEEE Computer Society, 2009: 23 –26.

[13] TRACE C B, DILLON A. The evolution of the finding aid in the United States-from physical to digital document genre[J]. Archival science, 2012, 12(4): 501 –519.

[14] SWALES J M. Aspects of article introductions[M]. Birmingham: the University of Aston in Birmingham, 1981.

[15] CROOKES G. Towards a validated analysis of scientific text structures[J]. Applied linguistics, 1986, 7(1): 57－70.

[16] HOPKINS A, DUDLEY-EVANS T. A genre-based investigation of the discussion sections in articles and dissertations[J]. English for specific purposes, 1988, 7(2): 113－121.

[17] SAMRAJ B. Introductions in research articles: variations across disciplines[J]. English for specific purposes, 2002, 21(1): 1－17.

[18] POSTEGUILLO S. The schematic structure of computer science research articles[J]. English for specific purposes, 1999, 18(2): 139－160.

[19] BRUCE I. Cognitive genre structures in methods sections of research articles: a corpus study[J]. Journal of English for academic purposes, 2008, 7(1): 38－54.

[20] BRETT P. A genre analysis of the results section of sociology articles[J]. English for specific purposes, 1994, 13(1): 47－59.

[21] KANOKSILAPATHAM B. Rhetorical structure of biochemistry research articles[J]. English for specific purposes, 2005,24(5): 269－292.

[22] NWOGU K N. The medical research paper: structure and functions [J]. English for specific purposes, 1997, 16(2):119－138.

[23] LEWIN B A, FINE J, YOUNG L. Expository discourse: a genre based approach to social science research texts[M]. London: Continuum, 2001.

[24] 赵福利. 英语电视新闻导语的语步结构分析[J]. 外语教学与研究,2001,33(2):99－104.

[25] 葛冬梅, 杨瑞英. 学术论文摘要的体裁分析[J]. 现代外语, 2005, 28(2): 138－146,219.

[26] 崔艳嫣, 王同顺. 英语学术讲座的宏观结构与微观结构——体裁分析在学术语篇分析中的应用[J]. 山东外语教学, 2004(5): 27－30.

[27] 杨瑞英. 体裁分析的应用:应用语言学学术文章结构分析[J]. 外语与外语教学,2006(10):29－34.

[28] BISHOP A P. Document structure and digital libraries: how researchers mobilize information in journal articles[J]. Information processing & management, 1999,35,(3): 255－279.

[29] DILLON A. Designing usable electronic text[M]. Boca Raton FL: CRC Press, 2004.

[30] DILLON A, SCHAAP D. Expertise and the perception of shape in information[J]. Journal of the American Society for Information Science and Technology, 1996, 47(10): 786－788.

[31] VAUGHAN M W. Identifying regularities in users' conceptions of information spaces: designing for structural genre conventions and mental representations of structure for Web-based newspapers[D]. Indiana: Indiana University, 1999.

[32] VAUGHAN M W, DILLON A. Learning the shape of information: a longitudinal study of Web-news reading[C]//Proceedings of the fifth ACM conference on digital libraries. New York:ACM, 2000: 236－237.

[33] DILLON A, SCHAAP D. Expertise and the perception of shape in information[J]. Journal of the American Society for Information Science and Technology, 1996, 47(10): 786－788.

[34] DILLON A. Spatial－Semantics: how users derive shape from information space[J]. Journal of the American Society for Information Science, 2000, 51(6): 521－528.

[35] ZHANG L, KOPAK L R, FREUND L, et al. A taxonomy of functional units for information use of scholarly journal articles[C]//Proceedings of the American Society for Information Science and Technology. MD, USA: American Society for Information Science Silver Springs, 2010,47(1): 1－10.

[36] ZHANG L, KOPAK L R, FREUND L, et al. Making functional units functional - the role of rhetorical structure in use of scholarly journal articles[J]. International journal of information management, 2011,31(1) : 21－29.

[37] ZHANG L. Grasping the structure of journal articles: utilizing the functions of information units[J]. Journal of the American Society for Information Science and Technology, 2012, 63(3): 469－480.

[38] MA C-C,CAO S-J. Identifying structural genre conventions across academic web documents for information use[C]//Proceedings of the Association for Information Science & Technology. Somerset, NJ: John Wiley & Sons, 2017: 260－267.

[39] 马雨萌, 刘凤红, 黄金霞. STKOS 中领域本体模型框架研究[J]. 图书情报工作, 2015. 59(3): 119－125, 139.

[40] 邱均平, 杨强, 楼雯. 资源本体构建理论与实证研究[J]. 情报理论与实践, 2014,37(5): 1－6.

[41] MIZOGUCHI R. YAMATO: Yet another more advanced top-level ontology[EB/OL]. [2019－10－28]. http://citeseerx. ist. psu. edu/viewdoc/download;jsessionid=00B4895D3EF153E0F74DC6B248D307FB? doi=10. 1. 1. 221. 1614&rep=rep1&type=pdf.

[42] 张囡囡. 面向语义网的领域本体半自动构建方法的研究[D]. 大连:大连海事大学,2008.

[43] GUARINO N. Semantic matching: formal ontological distinctions for information organization, extractiong and integration[C]//PAZIENZA M T. Information extraction: a multidisciplinary approach to an emerging information technology. Berlin: Springer Verlag, 1997: 139－170.

[44] 郭嘉琦. 领域本体的构建及其在信息检索中的应用研究[D]. 北京:北京邮电大学,2007.

[45] 王向前, 张宝隆, 李慧宗. 本体研究综述[J]. 情报杂志, 2016,35(6): 163－170.

[46] GRUBER T. Towards principles for the design of ontologies used for knowledge sharing[J]. International journal of human-computer studies, 1995, 43(5/6):907－928.

[47] 李景, 孟宪血, 苏晓路. 领域本体的构建方法与应用研究[M]. 北京:中国农业科学技术出版社,2009.

[48] DANIEL L R, NATALYA F N, MARK A M. Protégé: a tool for managing and using terminology in radiology applications[J]. Journal of digital imaging, 2007, 20(S1): 34 – 46.
[49] FREUND L. A cross-domain analysis of task and genre effects on perceptions of usefulness[J]. Information processing and management, 2013, 49(5): 1108 – 1121.
[50] 朱嘉贤,白伟华,李吉桂. Web 资源的多粒度语义标注及其应用技术研究[J]. 2011,38(8):83 – 87.
[51] 岳丽欣,刘文云. 国内外领域本体构建方法的比较研究[J]. 情报理论与实践,2016. 39(8): 119 – 125.
[52] LI Y, BELKIN N J. A faceted approach to conceptualizing tasks in information seeking [J]. Information processing and management, 2008, 44(6): 1822 – 1837.

作者贡献说明:
马翠嫦:开展研究的设计、实施和撰写论文;
曹树金:提出研究目标和研究思路。

Study on the Construction of Fine-grained Aggregation Ontology for Academic Documents in the Internet Environment

Ma Cuichang[1] Cao Shujin[2]
[1] Sun Yat-sen University Librariy, Guangzhou 510275
[2] School of Information Management, Sun Yat-sen University, Guangzhou 510006

Abstract: [**Purpose/significance**] Fine-grained information aggregation has become the focus in the field of knowledge organization. This paper aims at exploring the construction of fine-grained aggregation ontology for academic documents in the Internet environment. [**Method/process**] This study clarified the types, granularity characteristics and definitions of the concepts of the fine-grained aggregation ontology. Then, with the corpus of "citation analysis" documents in the field of library and information science in the Internet environment, the ontology was built through the concepts, attributes and relationships. At last, the ontology was evaluated and discussed. [**Result/conclusion**] This paper is among the first to propose the idea of the fine – grained aggregated ontology construction by using the concept of aggregation unit. This paper can inform the construction of knowledge organization systems for fine-grained organization, retrieval and navigation based on aggregation unit.

Keywords: information aggregation ontology academic document library and information science